

The evolution of a citation network topology:

The development of the journal Scientometrics

YIN LI-CHUN^{1,2}

HILDRUN KRETSCHMER^{1,3}

ROBERT A. HANNEMAN⁴

LIU ZE-YUAN^{1,2}

1. WISE LAB, Dalian University of Technology, Dalian, 116023, China E-mail:
yinliuxing1290@yahoo.com.cn.

2. Institute of Science Studies and Management of Science and Technology, Dalian University of Technology,
Dalian, 116023, China.

3. Department of Library and Information Science, Humboldt-University Berlin, Dorotheenstr. 26, D-10099
Berlin, Germany.

4. Department of Sociology, University of California - Riverside, USA.

Abstract

By mapping the electronic database containing all papers in Scientometrics for a 26-year period (1978-2004), we uncover the topological measures that characterize the network at a given moment, as well as the time evolution of these quantities. The citation network of the journal displays the characteristic features of a “small-world” network of local dense clusters of highly specialized literature. These clusters, however, are efficiently connected into a large single component by a small number of “hub” papers that allow short-distance connection among increasingly large numbers of papers. The patterns of evolution of the network toward this “small-world” are also explored.

Keywords: Citation network; Structure evolution; Scientometrics

1. Introduction

The statistical mechanics of complex networks has recently received considerable attention in several science communities, including statistical physics, computer science, and information science. The topological properties of large networks are a main focus of these studies. These

features of the topology of large complex networks are seen as important determinants of such processes as influence, diffusion, infection, and robustness. The real world networks that have been studied by physicists include the world wide web, the Internet (the physical connections between computers), email networks, the power grid of the United States, and numerous others [1].

Network topology can be applied to gain insights into the patterns of citation among scientific papers. Citation networks reveal patterns of influence on the development of new works, papers that play more central roles in the development of a literature, and the extent to which lines of inquiry form an integrated, cumulative body of scholarship. The properties of scientific citation networks have been studied in a number of papers[2_3_4]. Since Garfield's [5]pioneering work of on citation indexing for *Science*, and Price' [6] elaboration a decade later, the idea of analyzing networks among scientific citations has become widespread.

Citation studies can be used for many purposes, including the evaluation of the impact of individual scientists, papers, and institutions. Citation networks can also be used to understand the development of a scientific fields, and journals – as we do here. Our approach, unlike previous studies, focuses on the extent to which citation networks evolve toward “small-world” networks [1]. Small-world networks display certain topological features (high levels of clustering, short average path distances, and exponential degree distributions). These features are important characteristics of scientific communication patterns, because they indicate an overall coherence and integration across wide ranges of literatures, coupled with intensely connected local and specialized peer communities. Some previous studies have examined some of these aspects of the topology of citation networks. Redner [4] studied the citation distribution of 783,339 papers in journals cataloged by the *Institute for Scientific Information* and 24,269 papers published in *Physical Review D* between 1975 and 1994. He found that the probability that a paper is cited k times follows a power-law with exponent $\gamma_{\text{cite}} = 3$. Vazquez [3] extended these studies to the outgoing degree distribution as well, and found that it has an exponential tail. Lehmann et al. [2] studied the citation network in high-energy physics, and found similar distributions: for k less than 50, $\gamma_{\text{cite}} = 1.2$; for 50 or more citations, $\gamma_{\text{cite}} = 2.3$. The sociologist Hummon [7] studied the citation network in the field of DNA research. He emphasized the development of theory, and focused on locating a critical path through 40 milestone papers in DNA theory.

Scientometrics is one of the core journals in scientometrics field. As a representative journal of its field, it has been studied by many researchers [8,9,10,11] from different perspectives. This paper will study the *Scientometrics* citation network and its development with the view of citation network topology. By mapping the electronic database containing all papers in *Scientometrics* for 26 years period (1978-2004), we try to uncover the topological measures that characterize the network at a given moment, as well as the evolution of these quantities over time.

2. *Structure and evolution of citation networks*

The evolution of a journal in a specialized scientific field, such as scientometrics, is a lens on the formation of a scientific community. Seen as a network, earlier articles may become the targets of “ties” of citation from later papers[12]. The structure of a citation network could, in principle, take many forms. For example, later papers might not cite earlier papers, which would indicate a lack of coherence and cumulation in a field. Or, lines of inquiry may increasingly diverge over time, forming tightly connected “threads” that are disjoint from one

another. Or, lines of inquiry may become increasingly connected with time, as specialized inquiries become more aware of their commonalities, and begin to draw on more diverse literatures. As we examine the unfolding of the network with respect to time, we may find that certain papers act to cumulate the insights of those that came before.

Many network structures appear to display a tendency to evolve toward a “small-world” topology. The “small-world” network topology has several key features: clustering, short average path lengths, and an exponential degree distribution. Each of these features may be argued to have survival or fitness value in patterns of scientific communication. “Clustering” refers to the tendency of papers to cite the same other papers, forming dense local areas of the network that are display high mutual awareness. Successful scientific work requires comprehensive awareness of other relevant work, and fields progress through the intense competition and mutual support of narrow specialist communities. At the same time, however, it is important that specialist communities be aware of work of potential relevance in other specialist communities. A researcher working in one field, ideally, would also be “close” to work being done in all other specialist “clusters.” In the “small-world” each paper is primarily embedded in a narrow specialist literature, but is also at only a short “distance” from most other papers. This is achieved by the presence of proportionally small numbers of particularly “central” papers that cite multiple literatures (which gives rise to an exponential distribution of citations per paper). Successful papers of this type are widely cited (because they are relevant to many sub-fields), and form part of the “critical path” in the development of a field (because they cumulate prior work, and largely supercede it at the authority to which later work refers). The “small world” network, then, combines the features of intense specialization with easy access to relevant work in other specialites. It also connects all of the work in the diverse sub-specialties into a single over-arching “component” that defines the full field, and provides a “critical path” of cumulative development with respect to time. Because these features provide efficient, effective, robust, and innovative advantages, there should be selection pressures operating for the emergence of “small-worlds” in the literatures of scientific fields.

Below we will examine the extent to which the development of the literature in the journal *Scientometrics* resembles this ideal pattern. We begin by describing the data; we then examine the overall structure of the citation network as a whole (that is, for the entire period of 1978 to 2004); following this, we examine how certain features of the network have evolved with respect to time across this period.

3. Data

The first 59 volumes of the *Scientometrics*, were published between September 1978 and April 2004. These volumes contained 1853 items, which form the database of this study. Identification numbers have been assigned to the items sequentially, beginning with “1” (the first paper in volume 1) and ending with “1853” (the last item in volume 59). All types of items appearing in the journal are included. Table 1 shows the distribution of the items by type.

Table 1: Items appearing *Scientometrics* from September 1978 to April 2004, by type.

Type	Article	Meeting Abstract	Biographical-Item	Bibliography
Number	1527	11	13	17
Type	Note	Letter	Correction,Addition	Discussion
Number	74	16	10	1
Type	Review	Book Review	Item about Individual	Editorial Material

Number	30	89	8	57
--------	----	----	---	----

For our current purposes, we will examine only the citations among the 1853 items. That is, citations outside *Scientometrics* are excluded. This type of design precludes any conclusions regarding the relationships between *Scientometrics* and the larger scientific fields within which it is embedded. However, the “inner citation” approach is ideal for examining the structure of the discourse and interaction within the “context” of developing field of scientometrics [8].

A network is a pair $G = (V, E)$ consisting of two sets: a set of nodes $V = \{1, 2, \dots, N\}$, and a set of lines $E = \{e_1, e_2, \dots, e_L\}$ between pairs of nodes. If the line between two nodes is non-directional, then the network is called undirected network; otherwise, the network is called directed network. A network is usually represented by a graph, where nodes are drawn as small points, undirected lines are drawn as edges and directed lines as arcs connecting the corresponding two nodes. In citation networks, each paper is a node, and an arc (i.e., a directed line between two nodes) arises when one paper is cited by another, the head of the arc points to the cited paper.

Table 2: Types of items that are missing citation information.

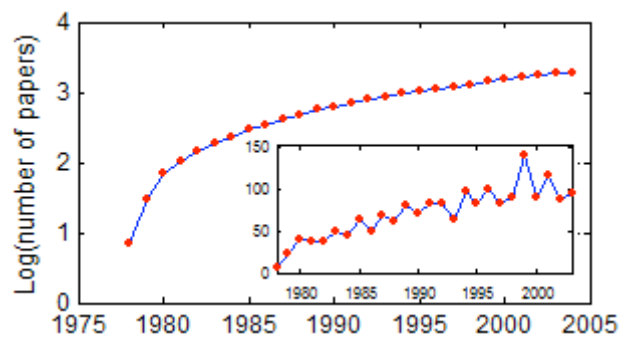
Type	Article	Review	Editorial Material	Bibliography	Item about Individual
Number	66	2	36	17	2
Type	Note	Letter	Meeting Abstract	Book Review	Biographical-Item
Number	13	4	10	1	5

The *Scientometrics* citation network contains 1853 nodes (items) and 5547 arcs (citations). Not all items, however, were ever cited: 442 items are “isolated” from all others within the journal. 157 of these “isolated” items appear to be possible data errors – for some reason, the database does not report citation data on them. The types of these “missing” items are reported in Table 2. The remaining 254 “isolated” items are true isolates – no other items in *Scientometrics* have ever cited them.

4. General characteristics of the *Scientometrics* citation network

Figure 1 shows the cumulative number of papers and the number of papers published each year during 26 years. The development trace of *Scientometrics* is very similar to that of other journals such as *JPHYS* and *HEP*[13]. The number of papers published each year is relatively constant, so the cumulative number of papers in network increase linearly with time.

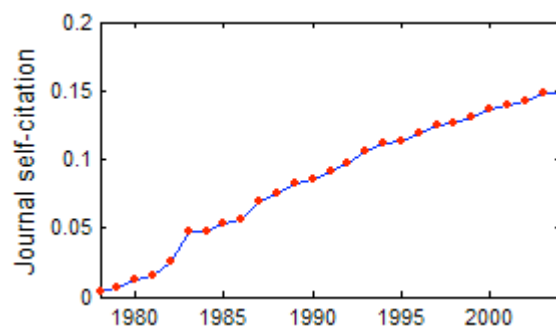
Fig. 1: Annual (inset) and cumulative number of papers in *Scientometrics* (1978-2004).



Ratios of “journal self-citation” and “journal self-cited” are two important indicators in journal evaluation. Both measures show the degree to which a journal has become recognized as the location in which important research is published.

The ratio of journal self-citation is the percentage of all references in papers published in the journal to other papers published in the journal. Figure 2 shows a steadily developing linear trend in self-citation in *Scientometrics* as the journal has aged.

Fig. 2: Proportion of citations in *Scientometrics* to earlier papers in the journal.

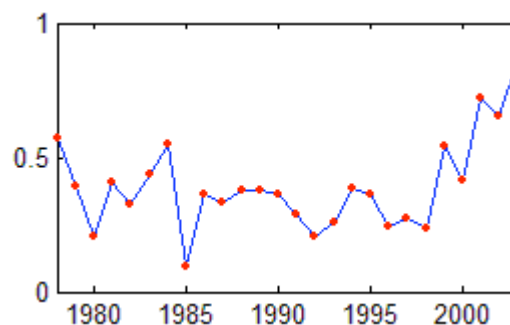


The ratio of “journal self-cited” is the percentage of all citations to papers in *Scientometrics* that are citations by other papers in the journal. This measure is available only for 2004, and stands at 63.5%. This rather high figure suggests that the journal has not yet achieved a highly central place in the entire literature cited by the SCI database.

Annual data on the ratio of “journal self-cited” are not available. However, we do have information on the proportion of papers in *Scientometrics* that have not been cited by papers in other journals. This trend is shown in figure 3.

The data for recent years must be discounted, as insufficient time has passed for the citation record to be complete. However, the earlier data show that *Scientometrics* has made slow progress in becoming fully integrated with the broader literature.

Fig. 3: Percentage of papers in *Scientometrics* not cited in other journals.

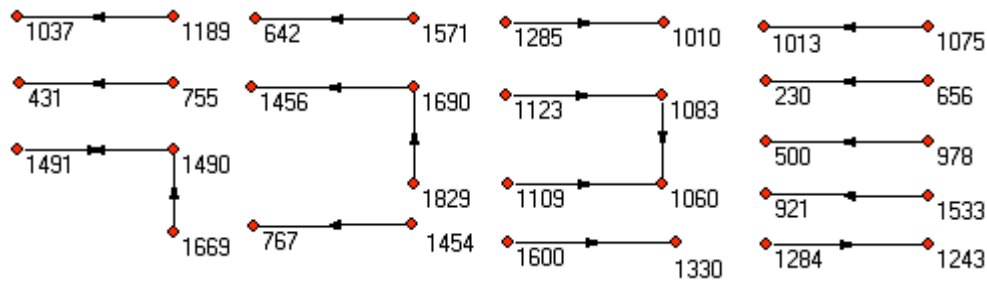


Components of a network are parts that are connected within, but disconnected between subgraphs [14]. This idea is readily interpretable in sociological terms. The members of a component can, in principle, communicate with one another, either directly or through chains of intermediaries. In citation networks, components identify how later contributions are connected to earlier ones. At one extreme, all later papers might be connected into a single integrated line of inquiry (i.e. all papers are connected in a single component); at the other extreme, there may be many separate sub-literatures. The pattern of components found in a network (their number and size) can, therefore, be taken as an indication of the opportunities and obstacles to communication or the transfer of ideas in the associated network [15].

In *Scientometrics* citation network, the largest component contains 1379 nodes. There are 14 rather small components, and 442 isolated nodes. From the distribution of components, we can see most of papers have some directed relationship with other papers within the biggest component. They either are the origins of some ideas, or receivers, or intermediaries. This suggests a rather high degree of coherence and cohesion in the literature published in *Scientometrics*.

The papers in 14 small components are disconnected with the main stream of knowledge. These components are shown in figure 4. These components of the research in *Scientometrics* have no bridge connecting them to the mainstream.

Fig. 4: Small components in the *Scientometrics* citation network.



One of the most striking features of the *Scientometrics* inner-citation network is the large number of papers (some 24%, or 411 nodes), that have no directed relations with other papers. Without considering self-citation, 44.1% papers have zero in-degree (are never cited) and 42.5% with zero out-degree (never cite another paper in the journal). Additionally, 95.17% of the papers in our network are not cited more than eight times. The top 4.83% papers contain more than 33% of all citations.

The median number of citations of papers in our database is 14.5, much larger than the mean citations (2.2). This suggests an extreme degree of skewness in the distribution. Using logarithmic scaling, an exponential relationship between the proportion of all citations (P) and the number of citations per paper (k) can be represented as a straight line. Figures 5 and 6 show the in-degree distribution (i.e. citations to papers) and out-degree (i.e. citations by papers) distributions.

Fig. 5: Distribution of in-degree in the *Scientometrics* citation network.

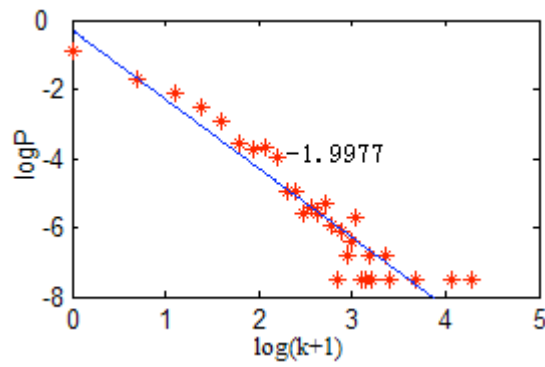
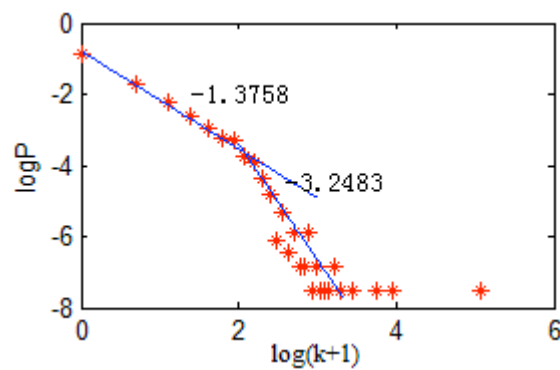


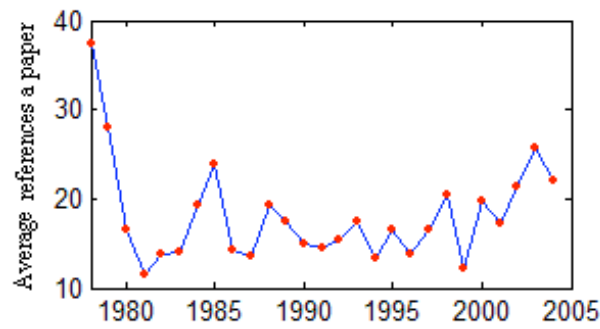
Fig.6: Distribution of out-degree in the *Scientometrics* citation network.



The probability that a random selected paper in the *Scientometrics* database has k citations or been cited k times are well described by simple power laws, $p(k+1) \sim k^{-\alpha}$, with $\alpha \approx 1.9977$ for in-degree. for out-degree distribution, the pattern is a little different, displaying a disjuncture at approximate 5 citations: when $k \leq 5$, $\alpha \approx 1.3758$, and $\alpha \approx 3.2483$ when $k > 5$. Both in-degree and out-degree distributions display clear evidence of “hubs” of central papers – both as sources, and as containers of citations. The overall degree of integration of the literature in *Scientometrics* is considerable, as single component connects a large proportion of all papers; and, key papers act as points that connect most papers at relatively short path distances.

5. *The evolution of Scientometrics citation network*

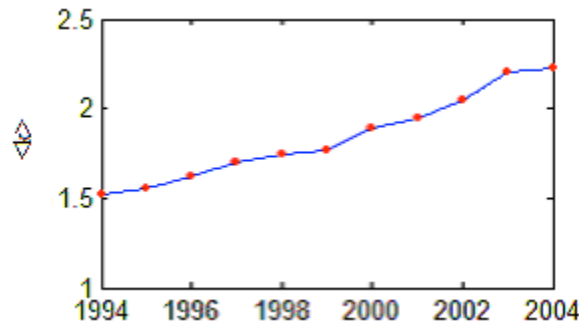
Fig.7: Average references to other papers in the journal.



If a journal is successful as a context for the publication of a cumulative scientific literature, the extent of internal citation should increase over time. Figure 7 shows that this is generally true of *Scientometrics*; the average number of citations to other papers in the journal has increased steadily over time after falling off for a high-peak of self-referencing in the founding volumes.

Similarly, if the research published in the journal is cumulative, the average amount of connection of subsequent to earlier works should increase with time. In figure 8, we see that this is the case; and that the trend toward increasing linkage to previous publications in the journal has increased in a rough linear pattern.

Fig.8: Average degree (ties to previous papers) in the *Scientometrics* database. Results are computed on the cumulative data up to the given year.



The majority of currently existing evolving network models assume a constant $\langle k \rangle$ as the network expands [16]. Such an assumption is consistent with the evolution of random graphs, or graphs displaying a tendency for recent new entrants to cite high-citation papers. The presence of a trend in the cumulative distribution of inner citation (figure 8) is indicative of a pattern of cumulative development, rather than a random connection. This pattern suggests a degree of coherence and cumulation that should characterize scientific (and other citation) data.

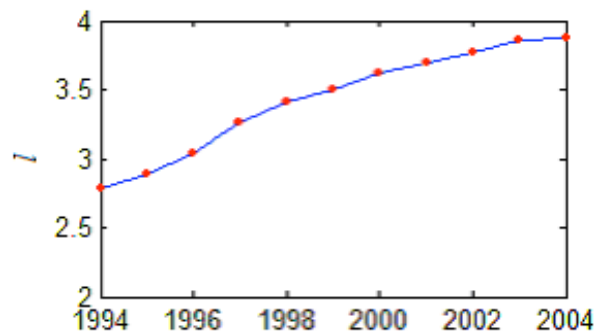
Nodes may be directly connected by a line, or indirectly connected through a sequence of lines. The length of a path is measured by the number of lines that make it up. The geodesic distance between two points is the length of the shortest path that connects them [15]. There may be many connections between two nodes in a network. If we consider how the relation between two nodes may provide each with opportunity and constraint, it may well be the case that not all of these path matter. The capability of two nodes, i and j , to transfer information largely depends on the length of the shortest path- d_{ij} - between them. That is, the geodesic path (or paths, as there can be more than one) is often the "optimal" or most "efficient" connection between two actors [14]. Mean geodesic distance is an indicator to evaluate the ability of information flowing in the whole network. Let l to be the mean geodesic distance between nodes pairs in a network:

$$\frac{1}{N} \sum_{j=1}^N d_{ij} \quad (1)$$

Where d_{ij} is the geodesic distance from node i to node j ; N is the total number of nodes in the connected component.

Figure 9 shows that the mean geodesic distance between papers in Scientometrics increases with time. However, the trend line has a shallow slope (the distance between papers does not increase by one unit for each additional volume added). And, importantly, the rate of increase in the mean geodesic distance decreases with time. This means that it is increasingly true that papers are equally close to all prior literature, even as the amount of prior literature continues to expand with the journal's development. This flattening of the path length is a key feature of the "small-world" effect.

Fig. 9: Mean geodesic distance in the *Scientometrics* database. The distance is computed on the cumulative data up to each year.



The *Scientometrics* citation network shows a clear small-world effect. "Small-worlds" are defined formally by Watts and Strogatz[17], and informally by Milgram [18]. It reflects the extent to which most pairs of nodes in most networks are connected by short paths. Milgram's experiment[18] showed a "six-degree separation" in American society. Our network displays "four degree separation" in 2004. In citation networks, small-worlds operate by way of "hub" papers that cumulate previous results, and allow later researchers to (indirectly) cite large streams of prior development by citing a "hub," rather than all of the previous pieces individually. That is, "hubs" allow later researchers to efficiently cite large literatures that are quite distant by citing, instead, a cumulative paper. This property makes the network quite efficient in terms of efficient complete searches.

In the small-world network, hubs act to make for short paths between large parts of the literature. The literature in scientific fields, is not random – but rather displays strong tendencies toward local clusters of narrow specialization. Most of the citations in most papers refer to quite narrow literatures directly relevant to a very specific topic; and all papers on that topic are linked directly to all others, forming dense local clusters.

Cliques form is a common property of social networks, representing circles of friends or acquaintances in which every member knows each other. In simple terms, the clustering coefficient of a node in the citation network tells us how likely it is that two of a node's references are to have citation relationship. It is easy to understand, when we think a paper is very 'useful', we are likely to stem from it to its citations, which made some of them have chances to be cited with this 'useful' paper together. then these papers are all the neighbors of our paper, and have citation relationship, vice versa. In this respect, clustering coefficient can be seen as the extension of co-citation and coupling in bibliometrics. This inherent tendency to cluster is quantified by clustering coefficient[19].

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \quad (2)$$

C_i is the ratio between the number E_i of links that actually exist between the neighbors of a selected node i , and the total number $ki(ki-1)/2$ of possible links between these neighbors. The clustering coefficient of the graph, which is a measure of the network's potential modularity, is the average over all vertices:

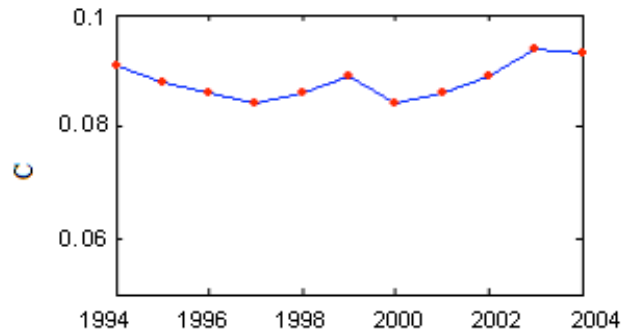
$$C = \frac{1}{N} \sum_{i=1}^N C_i \quad (3)$$

The range of C is $0 \leq C \leq 1$.

Figure 10 shows the development of the overall level of clustering in the citation graph over the history of the journal.

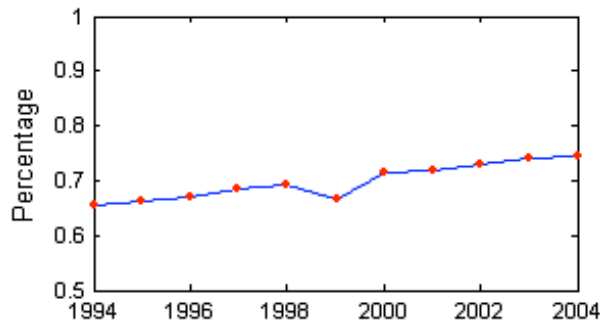
The level of clustering in the graph shows little overall trend, generally decreasing from 1994 through 1997, and generally increasing from 2000 to 2004. This is consistent “small-world” development, in that most of the citations in most papers are to papers that cite one another – that is, narrow specialties and sub-specialties. What is remarkable, however, is that the degree of clustering has remained strong while the overall amount of literature in the journal continues to increase; and while this literature is, overall, tied together at fairly short path-lengths.

Fig. 10: Clustering coefficient *Scientometrics* citations.



Knowledge diffusion processes can be looked at as adaptations and applications of knowledge documented in scientific publications. A closed connected citation network is important for transferring knowledge from one paper to another. If there are too many fragmented clusters in the network, knowledge diffusion will be set back seriously. Changes of the relative size of the largest component over time are shown in Fig.11.

Fig. 11: Percentage of all papers that are part of the largest component.



The relative size of the largest component increased continuously during the past decades. Up to April 2004, the scale of the largest cluster attained 1397 nodes, which is 75.4% of the whole database. Thus, despite the steady increase in the amount of the published literature, and despite the continuing tendency toward strong local “clustering,” the total body of literature in the journal is becoming more – rather than less – connected over time. Again, we see in this pattern the key importance of “hubs” that act as “shortcuts” in connecting large and increasing bodies of more specialized literatures into a large cumulative field.

6. Conclusion

Above we have combined elements from theories of complex network evolution, network visualization, and citation analysis in order to view the evolution of a journal in a new perspective. Focusing on the structure of citation network, we investigate the evolution of *Scientometrics* citation network with some network indicators. The input degree and output degree distributions in the *Scientometrics* citation network both show a scale free distributions. The high-citation “hubs” in the network act to integrate a very large part of all of the literature ever published in the journal into a single component, and create relatively short path lengths among clustered local specialty literatures, even as the literature increases steadily in size.

The existence of a “small-world” phenomenon in the citation network of *Scientometrics* is not surprising. The very logic of science, in a sense requires it. As scientific literature expands, it is not possible for all parts of it to be directly aware of all other parts of it; yet, if researchers in one specialty field are to benefit from advances in other fields, they must be connected. The emergence of “hubs” that cumulate knowledge and allow efficient search across diverse communities allows increasingly large and increasingly diverse and specialized literatures to remain connected – and provides a way for the entire scientific enterprise to be connected in a single large component – even though this emergent phenomenon is not apparent from the point of view of any one narrow specialty cluster.

Acknowledgement

This work greatly benefited from discussions with and comments from members of WISE-LAB of Dalian Technology of University, China. Borgatti, Everett, and Freeman’s Ucinet 6 program was used to calculate some measures. Borgatti’s NetDraw program was used to generate the network layouts.

References

1. Watts, D. J. (2003). *Six Degrees: The Science of a Connected Age*. New York: W. W. Norton.
2. S. Lehmann, B. Lautrup, and A. D. Jackson, (2003). Citation Networks in High Energy Physics. *Physical Review E*, 68, 026113
3. A. Vazquez, Statistics of citation networks, E-print: arXiv: cond-mat/0105031, 2001.
4. S. Redner, (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B*, 4 (2), 131-134.
5. E. Garfield, (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, 122, 108-111.
6. D. J. de Solla Price, (1965). Networks of Scientific Papers. *Science*, 149, 510-515.

7. NP.Hummon, and P.Doreian, (1989). Connectivity in a Citation Network: The Development of DNA Theory. *Social Networks*, 11, 39-63.
8. O.Persson, (2005). Exploring the analytical potential of comparing citing and cited source items. *Proceedings of ISSI*,1, 24-34.
9. B.Dutt, K. C. Garg, and A.Bali, (2003).Scientometrics of the international journal Scientometrics. *Scientometrics*, 56, 81-93.
10. 1A.Schubert, (2002).The Web of Scientometrics: A statistical overview of the first 50 volumes of the journal. *Scientometrics*, 53,3-20.
11. P.Wouters, L.Leydesdorff, (1994). Has Prices's dream come true: Is scientometrics a hard science? *Scientometrics*, 31, 193–222.
12. de Nooy, W., A. Mrvar, and V. Batagelj. (2005). Exploratory Social Network Analysis with Pajek. Cambridge: Cambridge University Press.
13. KB.Hajra and P.Sen, (2005). Aging in citation networks. *Physica A-Statistical Mechanics and ITS Applications*, 346 (1-2), 44-48.
14. R.A.Hanneman, (2004). Introduction to social network Methods. As e-print available at [http:// www.analytictech.com/networks.pdf](http://www.analytictech.com/networks.pdf) on May.25.2004.
15. J.Scott, (2000).Social network analysis-A handbook (2nd ed). Sage publications, London.
16. AL.Barabasi., H.Jeong, Z.Neda, et al. (2002). Evolution of the social network of scientific collaborations. *Physica A*, 311 (3-4), 590-614.
17. DJ.Watts, and S.H.Strogatz, Collective dynamics of 'small-world' networks. *NATURE* 393 (6684), 440-442, 1998.
18. S.Milgram, (1967). The small world problem. *Psychology Today*, 2, 60-67.
19. R.Albert and AL.Barabasi, (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74, 47-97.